5.2 Remote Users of SUMEX

Due to the fact that the SUMEX computer is available via both the TYMNET and ARPANET communication networks, it is possible for scientists in many parts of the world to directly access the Dendral programs on SUMEX. Primary usage is centered on CONGEN, although INTSUM is beginning also to gain a following. Although access points to SUMEX are widespread, they frequently are not diverse enough to accommodate the dispersed group of scientists who have expressed an interest in using one of the Dendral programs. For example, Dr. Joseph Baker of the Roche Institute of Marine Pharmacology in Dee Why, Australia, is looking at the possibility of accessing SUMEX by using International Direct Distance Dialing (IDDD).

5.3 Chemists Communicating by Mail

Many Scientists interested in using DENDRAL programs in their own work are not located near a network access point. Users of this type choose to use the mail to send details of their structure elucidation problem to a Dendral Project collaborator at Stanford.

5.4 Chemical Problems Posed to CONGEN

Following is a list of CONGEN users, and a brief summary of their program interests during the past year.

1. Dr. Roger Hahn, Syracuse University. While at Stanford he used CONGEN to help solve the structures of photoproducts by obtaining all possibilities under available constraints and designing NMR experiments to differentiate the possibilities. This work will be published soon.

2. Dr. William Epstein, University of Utah. During a demonstration of CONGEN, he posed a problem to verify that the structural possibilities he determined for an unknown were in fact all possibilities. The structure of methyl santolinate has been published (see Epstein, et al., J.C.S. Chem. Commun., 590 (1975)).

3. Dr. Clair Cheer, University of Rhode Island. While on sabbatical at Stanford, Dr. Cheer has worked on a number of structure elucidation problems using CONGEN including Briareine D and [+]-Palustrol (Cheer et al., Tetrahedron Letters, 1807 (1976)). Work is

continuing on the structure of another marine natural product, presumably a cembrenolide, for which there are currently seven possibilities.

4. Dr. Jerrold Karliner, Ciba-Geigy Corporation. Dr. Karliner has solved several structural problems using CONGEN, including materi . with flame retardant properties, an impurity in a production sample and nitrogen heterocycles being investigated for pharmacological activity. CONGEN enabled reduction of the number of possibilities to the point where subsequent experiments led to unambiguous structural assignment.

5. Dr. Gino Marco, Ciba-Geigy Corporation. He has used CONGEN to help solve structures of conjugates of pesticides with sugars and amino acids.

6. Dr. Milton Levenberg, Abbott Laboratories. He has worked on the structure of a compound with mild antibiotic activity, isolated from a fermentation broth. There are currently ten structural possibilities, reduced to that number from the 33 initially determined using CONGEN by additional experimental data.

7. Dr. David Pensak, DuPont. He is currently learning to use CONGEN and plans to evaluate its utility for structural problems of some of his coworkers.

8. Dr. Douglas Dorman, Eli-Lilly. He is using CONGEN to assist in structure elucidation of metabolites of microorganisms shown to have pharmacological activity. He has worked on five such problems, including a current one where the developing MSPRUNE capabilities are being used.

9. Dr. L. Minale, Napoli, Italy. We have worked with him by sending him structural alternatives for proposed structures for some marine natural products (Pallescensins, Tetrahedron Letters, 1417 (1975)) and cyclic diethers from the lipid fraction of a thermophilic bacterium (J. C. S. Chem. Commun., 543 (1974)).

10. Dr. K. Nakanishi, Columbia University. We have worked with him by sending him structural possibilities for termite defense compounds (structure finally solved by X-ray crystallography). This trial plus a live demonstration to one of his students has resulted in efforts toward continued collaboration on other insect defense secretions and

exploration of the possibility of his direct access to SUMEX.

11. Dr. L. Dunham, Zoecon Corporation. We have collaborated with him on the use of INTSUM for mass spectral fragmentation studies of insect juvenile hormones.

12. Dr. A. G. Gonzales, Tenerife, Spain. We have recently sent him structural alternatives for constituents of Laurencia Perforata (Tetrahedron Letters, 2499 (1975)), and expect to continue discussions on the structures of these compounds.

13. Dr. T. Irie, Sapporo Japan. We have recently sent him structural alternatives to published structures on constituents of Laurencia Glandulifera (Tetrahedron Letters, 821 (1974)) and expect to continue discussions on this problem.

14. Dr. C. J. Persoons, Delft. We have corresponded with him on structural alternatives for cockroach sex pheremones (Periplanone-B (Tetrahedron Letters, 2055 (1976)), and he has agreed to further collaboration on new problems.

15. Dr. F. Schmitz, University of Oklahoma. We explored for him structural alternatives for an unknown diterpenoid hydrocarbon. We obtained 25 possibilities, of which only four obeyed the isoprene rule.

16. Dr. J. Baker, Roche Institute of Marine Pharmacology, Australia. We plan collaboration with Dr. Baker on the sterol fractions of various marine organisms and are exploring ways for him to access CONGEN.

17. Dr. E. VanTamelen, Stanford University. We have used the developing reaction features of CONGEN to explore structural possibilities for both chemical and biogenetic cyclization products of squalene-oxide congeners. We have suggested alternatives to proposed structures and helped to design experiments to differentiate them.

18. Dr. J. C. Braekman, Brussels. Dr. Braekman visited Stanford as a part of continuing collaboration in marine chemistry with Dr. Tursch's group. While at Stanford he explored use of CONGEN for use in current problems in marine natural products, and worked on the problems of Drs. Irie and Gonzales (see above).

He is currently exploring access to CONGEN from Brussels, via TYMNET.

Some problems have arisen as a result of the Dendral commitment to working with outside chemist users. The primary area of difficulty arises from the fact that the Dendral project, as one of the many projects which use the SUMEX facility, is allocated a certain portion of system resources. Therefore, support of an extensive body of outside users means that resources to support these users must be diverted from the research goals of the project.

In encouraging new users, Dendral must be careful to state that access to Dendral programs might have to be restricted in the future if system loading becomes extensive. Understandably then, some scientists are reluctant to invest time in learning to use a complicated, although potentially useful program which they may well only be able to use on a temporary basis. One solution to this problem is to make the available programs as efficient as possible, and/or to make it possible to distribute copies of the program to other sites.

Use of CONGEN by working scientists has turned up one major area in which additional information to the user was thought to be necessary. CONGEN users unanimously indicated their desire for a method what percentage of the whole problem was solved at any moment, i.e., total number of possible structures is represented by the number already generated. In a prototype system we have implemented the Cntrl-I and Cntrl-S user information interrupts, to show how far CONGEN has progressed. If, for example, someone who has generated 357 structures is told that this indicates that they have generated 1 percent of the total possible structures, they immediately know that they do not want to finish generating all the structures. Even if there were enough space, 40,000 structures would be far more than they would want to see.

We implemented another user-oriented facility for an invited paper presented at the 172nd American Chemical Society meeting, in August of 1976. Special features were added for a character-oriented, screen-addressable CRT terminals to give users an informative visual interface to CONGEN, an otherwise complex The dynamic field of view provided by this type of terminal was used to advantage to give the chemist-user a continuous, graphic summary of both the information he has supplied to the program and the dynamic use of that information by the program.

6 Stereochemistry in CONGEN

We have started the complex task of giving CONGEN the capability of recognizing stereochemical features of molecules and using stereochemical information in structure determination. The ability to recognize stereochemical features would allow, for example, the generation of all stereoisomers of a given topological structure with or without constraints. The ability to use stereochemical information would allow the determination of constraints on stereoisomer (and topological isomer) generation caused by, for example, partial knowledge of relative or absolute stereochemistry of structural fragments, knowledge of overall molecular chirality (or lack of), absolute and relative stereochemistry from circular dichroism measurements, and so forth. Thus far, only the topological information (constitution) has been recognized and used by CONGEN.

The first stage of this development is to produce a program which generates all the stereoisomers of a given topological structure. This program will be placed at the end of the existing CONGEN program. The present report describes the development of the theory and algorithm for stereoisomer generation and the progress on the programming of this algorithm.

6.1 Algorithm

The carbon stereoisomers of a given topological structure are in correspondence with the double cosets:

TSG[A4] / TSG[S4] / CSG

in which:

1. TSG[A4] is the wreath product of the Topological Symmetry Group and the alternating group A4. This group expresses the invariance of a carbon stereoisomer to all even permutations of the ligands connected to any carbon stereocenter.

2. TSG[S4] is the wreath product of TSG and the symmetric group S4. This group expresses the invariance of the connectivity of a topological structure to all permutations of ligands connected to any carbon center.

3. CSG is the Configurational Symmetry Group and is isomorphic to the TSG represented on the two-valued configurations of the carbon stereocenters.

The cosets of TSG[A4] in TSG[S4] correspond to the 2^m maximum possible stereoisomers where m is the number of carbon

stereocenters. The effect of the group CSG on these cosets is to collect the possible stereoisomers into equivalence classes of distinct stereoisomers Intuitively this corresponds to the mental process of considering all possible stereoisomers of a topological structure and collecting those equivalent by symmetry.

The algorithm to generate stereoisomers from a CONGEN topological structure must perform three transformations:

1. The connection table (CT) corresponding to the CONGEN topological structure must be modified to include only those carbon centers which need be considered as stereocenters. That is, methylenes, methyls, carbons with gem-dimethyls etc., do not exhibit configurational stereochemistry. A prefilter must act on the CT and return a Stereocenter Connection Table (SCT).

2. The TSG which comes from CONGEN must be modified to give the CSG described above.

3. Given the SCT and CSG, the possible distinct stereoisomers must be generated. This involves an implementation of the theory presented in the previous paragraphs. Further details of this algorithm are given in the next section.

6.2 Programming Progress

All programming is being done in the SAIL language.

1. The development of a program to perform the prefilter function on the connection table is currently in progress. The CT will first be scanned to eliminate methylenes and methyls and then iteratively scanned to find identical achiral substituents on common carbons (gem-dimethyl, gem-diethyl, etc.).

2. A program to obtain the configurational symmetry group (CSG) from the topological symmetry group (TSG) has been written. The elements of TSG are allowed to act on the connection table and the parity of the permutations on each stereocenter is determined. The permutation with these parity designations is the desired element of CSG.

3. A program which, when given the SCT and CSG, will generate all distinct stereoisomers has been written. Special use is made of the fact that all elements of the CSG will be hyperoctahedral group elements. That is, CSG will be a subgroup of the wreath product Sn[S2], called the hyperoctahedral group, where n is the number of stereocenters. The order 2 group, S2, is represented by the two-valued configuration of each carbon stereocenter. This two-valued nature of each stereocenter's

configuration is easily represented by a single two-valued bit which makes a very compact machine representation. The program has the capability of representing the hyperoctahedral group by bit permutation and reversals. This will accommodate any conceivable symmetry and any stereochemistry resulting from carbon (or analogous element) configurations.

As an example, consider the problem of the number of stereoisomers of inositol, (CH(OH))6. The CSG can be obtained from the TSG as described above and when input with the stereochemical connection table to this segment of the program, the desired 9 isomers are found and output as canonical structures based on the original atom numbering. (This will probably not be the final choice for a canonical stereostructure.) The interfacing of these segments of the stereoisomer generator and the interfacing with the existing CONGEN program is also in progress.

7 The GC/HRMS DATA SYSTEM

7.1 Improvements to the Data System

The introduction of the gas chromatograph (GC) into the high resolution mass spectrometry (HRMS) system produced a number of problems in data reduction that are not present without the GC. The primary problem is the increase in the number of mass peaks in a spectrum from the column bleed of the GC. This makes the problem of separating calibration and reference peaks from the true sample peaks a much more difficult problem. A number of improvements have been introduced to the software to solve this problem.

The instrument is calibrated by injecting a sample of perfluorokerosene (PFK) and running REFRUN. This collects a spectrum which can be calibrated by looking for various characteristic peaks in the spectrum. The masses of certain peaks are stored on a file. Once these calibration peaks have been identified, the masses can be used to interpolate and find the mass of all other peaks in the spectrum. The results of a satisfactory reference run is stored on a file, as well as being listed on a line printer.

The spectrum of the sample is taken by runing SAMRUN, which collects a spectrum of the sample and PFK. The main problem now is finding the peaks from PFK, and using them to calculate the masses of the peaks from the sample. The first ten calibration peaks are located by applying a template, or pattern matching algorithm to the data. This template assumes that characteristics of the mass spec will change only systematically with time. This has proven to be a very successful and sensitive method of locating calibration peaks. Once the initial ten peaks are located, the program scans the data by taking four calibration peaks and, using a model of the scan, projecting for a fifth. Once this is located the masses of the peaks in between the calibration peaks are interpolated, and a decision is made on whether a given peak was in the reference run, or is truly a sample peak. The four calibration masses are shifted so that the calibration peak just projected becomes one of the four, and the process is repeated until masses have been assigned to all of the peaks.

Problems occur when, during projection, either no peak or more than one peak is found as a calibration peak. If no peaks are found, the mass is counted as missing, and the next calibration mass is searched for. Since the calibration peaks are chosen as being among the most prominent peaks in the spectrum, the problem in this case is usually not that the peak is absent. The more common problem is that there are so many data peaks from the GC that more than one peak shows up as a candidate for the calibration peak. If the program chooses the wrong peak as the calibration peak, the crawl through the data quickly goes bad. Various schemes have been tried to minimize this problem. Originaly the first peak in the window was chosen, since PFK has a very large negative mass defect. This produced occasional problems, however. Next a more sophisticated approach was tried. If the projection produced multiple candidates for the calibration peak, the two peaks closest to the projection were selected, and another projection was done from each of those. The one giving rise to the least total projection error was selected. We found one batch of data, however where it happened that at one section of the spectrum, two incorrect peaks produced a total error less than the two correct peaks. Neither of these algorithms use any information from the reference run, so an attempt was made to fold in the information from the reference run in the case of an ambiguity.

The spectrum is taken in an exponential downscan, i.e. high mass to low mass on an exponential curve. The only two parameters of this curve that can change are the time offset of the curve and the time constant of the exponential. The template mentioned earlier assumes that either of these parameters can change and attempts to find a set of peaks in the sample run that map most accurately into the reference run. This mechanism works well only in low masses, however, since in the higher masses the

curve is more gaussian than exponential. The template can be written for this, but the amount of space and time required for it made it appear impractical for the system. On examination of data it became obvious that the time constant of the exponential changed very little, if at all, from the reference run to the sample run. This means that a very good approximation of where a given calibration peak should appear can be obtained by merely adding in the time shift from the reference run. The final algorithm that resulted goes through the following steps: 1) the fifth calibration peak is projected. 2) if there is more than one candidate, then a projection is done on the two closest calibration peaks. 3) if both of these peaks project to another peak (there is still ambiguity, in other words), the peak which is closest to the time in the reference run based an exponentially weighted time shift from the previous calibration peaks is chosen. This has proven to be fast and reliable on the data tested so far, including data that had produced incorrect results from the previous two methods.

7.2 Changes in the Operating System

The current operating system for the PDP-11, DOS 9, has produced a number of problems. Poor keyboard interaction, generally slow response time, and extremely slow system programs, while surmountable, are factors that make the system difficult to use. We decided to look at the feasibility of changing to either RSX-11M or RT-11. RSX-11 proved to be too big and much more flexible than needed. RT-11 however had several advantages over DOS 9. The keyboard interaction is easier to use and more suited to a real time environment. The IO queuing structure is much simpler and faster, although the file structure is not as flexible as DOS 9. In addition the system itself is much faster, and there is a noticable improvement in time of just loading programs. Based on this, a decision was made to switch from DOS to RT-11.

The conversion of programs from DOS to RT-11 has proven to be much more work than originaly expected. The main problems have been incompatibilities between the two versions of FORTRAN and the different linkage editors. Since all the programs in the high resolution system are overlayed, this second factor has proven to a major problem, since some of the logic in the program must be reworked to make the overlay correct. The conversion effort has been aided by several factors, however. The speed of the system and system programs is often several times faster than similar programs under DOS. As an example, to link the REFRUN portion of the High Resolution system takes about 30 minutes under DOS, whereas the same program takes about 5 minutes under RT-11. The FORTRAN compiler and MACRO assembler are also faster.

The conversion, and software development in general, has been greatly improved by the addition of a teletype line from the SUMEX PDP-10 to the PDP-11. Programs have been writen to transfer files between the two systems. This has had the effect of switching literaly all of the editing to SUMEX because of the superior editor. The ease and speed of the file transfer makes it practical to make even minor modifications of a program on the the 10, and then transfer the edited version to the PDP-11. This process of using SUMEX to evelop software will continue with the release of MAINSAIL, a machine independent language. MAINSAIL is a dialect of SAIL, which is a dialect of ALGOL 60. It has undergone many design changes since its original inception, but has been released in a limited version for the PDP-10. The main value of MAINSAIL is that programs written on one machine will be directly transportable to another with no modification. This allows us to write, test and debug software systems on SUMEX, which leaves only the the machine dependent portion of the system (for example the actual real time data acquisition) to be worked out on the PDP-11. This not only gives the programmer better tools (such as superior editors) but also frees up the PDP-11 for production work.

7.3 New Developments

In addition to upgrading old versions of the high resolution system, work is being done on creating a low resolution system for the MAT 711. The ultimate aim is collect data that can be run through CLEANUP, a program that resolves multiple spectra under a single GC peak, and cleans up the final spectra. The problem with the current system is that we cannot scan fast enough to provide CLEANUP the data it needs. The high resolution system requires resolution good enough to separate sample peaks from the reference peaks. If the scan is sped up past a certain point, SAMRUN can no longer separate the peaks, and therefore cannot calibrate the run. At the same time, CLEANUP requires at least 7 spectra across a GC peak be taken to insure resolution of multiple spectra. The fundamental problem then is that an alternate method of calibrating the mass spectrum, without using known calibration peaks, must be found before scan speeds required by CLEANUP can be achieved. The most direct solution to this is to directly measure the magnetic field strength of the instrument, and using it to calculate the mass that is being observed. To do this we inserted a hall probe between the poles of the magnet, and connected it to the data acquisition system on the PDP-11/20.

The main problems with the hall probe are as follows: 1) to make sure that the ion reading and the hall probe reading are simultaneous 2) to insure that the correct hall reading can be assigned to the correct ion reading 3) to determine the

reproducibility of hall readings versus mass being observed in both dynamic (scanning) and static situations and 4) to decide if the probe has the speed and accuracy to calibrate the instrument. The first two problems are a matter of hardware. The configuration of the original data collection system is as follows: the ion detector goes to an A/D converter, which is connected to a DMA. The DMA is on an 11/20, which has a data collection system, SAQMON, running. This performs various low level filtering and buffering operations. The DMA is actually a low level processor which counts the number of samples taken, stores them into successive memory locations, and interrupts the central processor when a block of data has been collected. The timing of the sample collection is controled by a quartz crystal clock. On each timing pulse, a signal is sent to the A/D on the ion detector to convert that value to a digital number. To accommodate the hall probe, the DMA was modified so that on the timing pulse, the start signal is sent simultaneously to both the A/D on the ion detector and the A/D on the hall probe. The DMA then services both of the A/D's, and stores the readings in successive memory locations. The net result is that when the DMA interrupts the central processor, the block of data is a set of pairs of readings, an ion reading and the hall reading for that time. This solves both of the first two problems, since we now have the ion reading and the hall reading connected both in time and location.

The second two problems, testing the reliability and reproducibility of the hall probe, requires new software. We are currently modifying portions of the calibration mechanism of the high resolution system to calculate masses for a large number of hall readings.

8 META DENDRAL

The success of any reasoning program is strongly dependent on the amount of domain-specific knowledge it contains. This is now almost universally accepted within AI, partly because of DENDRAL's success. Because of the difficulty of extracting specific knowledge from experts to put into the program, many years ago we began to explore the problems of efficiently transferring knowledge into a program. We have looked at two alternatives to "hand-crafting" each new knowledge base: interactive knowledge transfer programs and automatic theory

formation programs. In this enterprise the separation of domain-specific knowledge from the computer programs themselves has been a critical component of our success.

One of the stumbling blocks with the interactive knowledge transfer programs is that for some domains there are no experts with enough specific knowledge to make a high performance problem solving program. We were looking for ways to avoid forcing an expert to focus on original data in order to codify the rules explaining those data because that is such a time-consuming process. Therefore we began working on an automatic rule formation program (called Meta-DENDRAL) that examines the original data itself in order to discover the inference rules for that part of the domain.

The problem solving paradigm for Meta-DENDRAL is also the plan-generate-test paradigm used in Heuristic DENDRAL. In this case one part of the program (RULEGEN) generates plausible rules within syntactic and semantic constraints and within desired limits of evidential support. The model used to guide the generation of rules is particularly important since the space of rules is enormous. The planning part of the program (INTSUM) collects and summarizes the evidential support. The testing part (RULEMOD) looks for counterexamples to rules and makes modifications to the rules in order to increase their generality and simplicity and to decrease the total number of rules.

Meta-DENDRAL successfully formulated rules of mass spectrometry that were new to the science. These rules, along with a discussion of the methodology, were published in the scientific literature [Report HPP-76-4]. The program was tested to see if it could rediscover the rules of mass spectrometry for two classes of chemical compounds that were already well understood (amines and estrogenic steroids). Then it was applied to three classes of compounds whose mass spectrometry was not as well known (mono-, di-, and tri-ketoandrostanes). The program produced three sets of rules that explained much of the significant data for these classes. The time for manual rule formation for these data was estimated to be several months.

Progress was made on generalizing the Meta-DENDRAL program, and rules for a new domain were successfully discovered by the program. A scientific paper on this application was submitted for publication [Report HPP-77-4]. The new application was learning rules for interpreting signals from C13-NMR spectroscopy. The instrument produces data points in a bar graph in response to the resonance of each carbon-13 nucleus in the sample. The rules describe an environment of a C13 atom and predict a resonating frequency range for every atom that matches the description. The Meta-DENDRAL program needed some modification because the rules are predicting ranges of data points, and not precise processes, as for the mass spectrometry version.

The RULEGEN component of Meta-DENDRAL was demonstrated to work with its heuristic search paradigm. Guidance from a model of mass spectrometry is an important feature of RULEGEN. Also, the program uses problem data for pruning possible rules (and all more specific rules formed from those). The amount of data examined during the search is very large and the space of rules is immense, so the search needs to be rather coarse in order to produce plausible, but not necessarily optimal, rules.

The RULEMOD program for "fine-tuning" Meta-DENDRAL's newly-discovered rules was finished. This program provides a number of important subtasks, including merging similar rules, making rules more specific or more general, and filtering out the weakest rules. RULEMOD checks for counterexamples to rules and uses this information in all of the named tasks. Because of the expense of computing counterexamples to possible rules, this computation is delayed until Meta-DENDRAL has a set of plausible rules, rather than computing counterexamples on each possible rule examined in the search of the rule space.

A report was written on the AI methodology underlying Meta-DENDRAL The major idea developed in this report is that knowledge of the domain can be used effectively to guide a learning program. The major difference between Meta-DENDRAL and statistical learning programs is that Meta-DENDRAL uses a strong model of mass spectrometry, including any assumptions the user cares to make about the domain, to guide the formation of explanatory rules.

9 C13 NMR SPECTROMETRY

13C NMR was selected as a new application area for the rule formation program, Meta-DENDRAL. The algorithms used for mass spectrometry rule formation were extended to 13C NMR and used to obtain a set of rules for These two classes and acyclic amines. These two classes were chosen since compounds in these classes are known to show a strong correlation between structural environment and shift. Thus, the programs could be tested knowing that the underlying basis for the form of the rule was valid.

The form of the rule is

substructure ---> shift range.

A sample rule generated is

C-C*-C-X- ---> 19.85<= (delta sub C)<=21.3.

The asterisk in the substructure description denotes the atom for which the shift is predicted. Only topological descriptors were used to construct the substructures. The addition of stereochemical terms is a topic of current work.

It was necessary to change RULEGEN so that the left-hand sides of rules were expanded outward from a carbon atom rather than from a bond. The right-hand side of the rule is associated with a range rather than a precise mass as in the mass spectrometry program. This modification also required changes in the rule search procedure. The user sets two parameters which guide the rule search. These parameters are MINIMUM-EXAMPLES which requires each rule to explain a given number of peaks in the training set and MAXIMUM-RANGE which defines the acceptable shift range for a rule. These parameters regulate the degree of specificity or generality of the rules.

From the set of rules generated a subset is selected corresponding to the "best" set which still covers all the training set data. The best rule is selected by calculating

(number of peaks predicted/(range ** 2)).

Data which are predicted by the best rule are removed and the next best rule is found for the remaining data using the criterion given above. This process is repeated until all data are explained.

In order to test the informational content of the rules generated a second program was written which applied the rules to a list of candidate molecules and ranked the molecules. Firsts, all possible structural isomers for a given empirical formula were generated using CONGEN. The rules were applied to each of the possible isomers and spectra were predicted. The predicted spectra were compared to that of a known spectrum from a compound with the same empirical formula. The structural isomers were ranked according a comparison score to determine how well the correct compound was distinguished from its isomers, on the basis of the predictive rules.

The details of the generation of rules and the use of rules for structure selection can be found in a paper recently submitted for publication [Report HPP-77-4]

The 13C NMR rule formation program was applied to a set of paraffins and acyclic amines. The program generated 138 rules to cover 435 data peaks. The rules generated were applied in a structure selection test for the structural isomers of C9H20 and C6H15N. No structures with these empirical formulas were

included in the training set. Twenty-four C9H20 and eleven C6H15N 13C NMR spectra were available to act as unknowns in the structure selection test. The results of the structure ranking applied to these spectra are shown below.

EMPIRICAL FORMULA	NUMBER OF CANDIDATE ISOMERS	NUMBER OF CANDIDATES RANKING			
		1st	2nd.....6th.....9th		
C9H20	35	20/24	3/24		1/24
C6H15N	39	8/11	2/11	1/11	

The performance of the rules in discriminating among similar structures not included in the training set data demonstrated the content of the rules.

10 BUDGET

Budget Information relevant to future funding was submitted with the renewal proposal to the BRP.

11 RECENT PUBLICATIONS OF THE HEURISTIC PROGRAMMING PROJECT

(Only publications related to computers in chemistry are shown.)

HPP-76-1 D.H. Smith, J.P. Konopelski and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures", Organic Mass Spectrometry, 11: 86, (1976).

HPP-76-2 Raymond E. Carhart and Dennis H. Smith, "Applications of Artificial Intelligence for Chemical Inference XX. Intelligent Use of Constraints in Computer-Assisted Structure Elucidation", Computers In Chemistry (in press).

HPP-76-3 C.J. Cheer, D.H. Smith, C. Djerassi B. Tursch, J.C. Braekman and D. Daloze, "Applications of Artificial Intelligence for Chemical Inference XXI. Chemical Studies of Marine Interbrates - XVII. The Computer-Assisted Identification of [+]-Palustrol in the Marine Organism Cespitularia sp., aff. subviridis". Tetrahedron. 32:1807, Pergamon Press, (1976).

HPP-76-4 B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg, and Carl Djerassi, "Application of Artificial Intelligence for Chemical Inference XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program", Journal of the American Chemical Society, 98: 6168 (1976).

HPP-76-5 T.H. Varkony, R.E. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference XXIII. Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems", in "Computer-Assisted Organic Synthesis", W.T. Wipke, Ed., American Chemical Society, Washington, D.C., in press.

HPP-76-6 D.H. Smith and R.E. Carhart "Applications of Artificial Intelligence for Chemical Inference XXIV. Structural Isomerism of Mono and Sesquiterpenoid Skeletons 1,2-", Tetrahedron, 32:2513, Pergamon Press (May 1976).

HPP-76-10 Bruce G. Buchanan and Dennis Smith, "Computer Assisted Chemical Reasoning", in Proceedings of the III International Conference on Computers in Chemical Research, Education and Technology", Plenum Publishing, (1976).

HPP-77-4 T.M. Mitchell and G.M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference. XXV. A Computer Program For Automated Empirical 13C NMR Rule Formation", (Submitted to JACS, January 1977).

HPP-77-6 STAN-CS-77-597 Bruce G. Buchanan and Tom Mitchell. "Model-Directed Learning of Production Rules", Submitted to the Proceedings for the Workshop on Pattern-Directed Inference Systems in Hawaii, (February, 1977).

HPP-77-11 Dennis H. Smith and Raymond E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data". Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry. University of Nebraska, Lincoln, (in press).